Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios

Porcel, Eduardo; Dapozo, Gladys; López, María V.
Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura
Universidad Nacional del Nordeste. 9 de Julio Nº 1449. CP 3400. Corrientes. Argentina.
TE: (03783) 423126 - (03783) 473930 Fax)
{gndapozo, eporcel, mvlopez}@exa.unne.edu.ar

CONTEXTO

Las líneas de I/D presentadas en este trabajo forman parte de las actividades definidas en el marco del proyecto F008-2008: "Rendimiento Académico de alumnos de la FACENA – UNNE: Su análisis mediante métodos cuantitativos", acreditado por la Secretaría General de Ciencia y Técnica de la Universidad Nacional del Nordeste. El mencionado proyecto tiene como objetivo fundamental construir modelos predictivos del rendimiento académico de los alumnos de las carreras de la Facultad de Ciencias Exactas y Naturales y Agrimensura (FACENA). A tal fin, puede ser considerado una extensión del proyecto de investigación (PI 005/06) anteriormente desarrollado por el mismo grupo de investigación.

Los resultados obtenidos en el mencionado proyecto, han permitido determinar, principalmente, la estrecha vinculación que existe entre el rendimiento académico de los alumnos del primer año de todas las carreras con el nivel de conocimientos matemáticos previos y con las condiciones socioeconómicas de los mismos. Así también, para las carreras de formación docente y licenciaturas en ciencias básicas, se pudo observar que el escaso nivel de avance en los estudios y las prolongadas estadías de estos alumnos en el sistema guardan relación con la estructura curricular, fundamentalmente con el ordenamiento de los contenidos de enseñanza (correlatividades).

Sin embargo, por la permanente mutación que sufren los factores mencionados (fluctuaciones en las condiciones sociales y económicas de los alumnos, desarrollo de programas de mejora de la calidad de la enseñanza en los niveles educativos previos, reformas de los planes de estudios, entre otros), es factible pensar que las variables correspondientes a estas dimensiones, no permanezcan estables con el devenir del tiempo y, por lo tanto puedan ser reconstruidas las relaciones y tipologías obtenidas.

De acuerdo a lo expresado anteriormente, en las líneas de investigación presentadas en este trabajo, se ha puesto énfasis en la construcción de modelos matemáticos que permitan predecir el rendimiento académico futuro de los estudiantes, tomando como base la información de las cohortes 2001 – 2008. Esta predicción permitirá conocer el rendimiento académico del alumno a priori con sólo disponer de la información de los mismos referida a sus condiciones iniciales (socioeconómicas y/o de conocimientos matemáticos previos), detectar con anticipación cuáles son las acciones pertinentes para contribuir a que los estudiantes superen los obstáculos que actualmente les impide avanzar en sus estudios y finalizarlos en menos tiempo que el que hoy día emplean.

RESUMEN

Este proyecto tiene por objetivo construir modelos predictivos del rendimiento académico de los estudiantes de las diversas carreras de la FACENA de la UNNE. Las variables a incorporar en los modelos serán seleccionadas de acuerdo a los resultados obtenidos a partir de los siguientes análisis: a) Resultados del test de diagnóstico de conocimientos matemáticos previos; b) Condiciones socioeconómicas de los alumnos de las distintas carreras y datos obtenidos de encuesta directa a los alumnos de primer año. Para la formulación y ajustes de los modelos de predicción, se utilizarán alternativamente, técnicas de minería de datos clásicas y métodos simbólicos o inteligentes, evaluando su desempeño en la predicción del rendimiento académico de los alumnos. Los resultados obtenidos a partir del desarrollo de este proyecto, constituirán un aporte significativo para los procesos de evaluación y acreditación universitarios, considerando que la reflexión sobre todos los elementos proporcionados por el análisis del rendimiento del alumnado contribuirá a la mejora de la calidad del sistema educativo.

Palabras clave: *Rendimiento académico. Educación Superior. Minería de datos. Métodos simbólicos. Métodos estadísticos.*

1. INTRODUCCION

A partir de la década del '80 surge en las universidades de todo el mundo la preocupación por la calidad del servicio educativo que prestan. Esto dio lugar a procesos de evaluación a fin de detectar las debilidades y fortalezas institucionales y generar acciones correctivas de las deficiencias encontradas.
En nuestro país, en la década del '90, el Estado Nacional incluye en su agenda de política educativa

la evaluación de la calidad del accionar universitario, y la mayoría de las universidades nacionales inician procesos de evaluación institucional.

En 1996, se conocen los primeros resultados referidos al rendimiento académico de los estudiantes de las trece carreras que por entonces podían cursarse en la FACENA. Dicha información hace referencia a elevados índices de desgranamiento en todos los años de estudios pero, fundamentalmente, al término del primer cuatrimestre del primer año de estudios. Asimismo, da cuentas de que el retraso promedio en el egreso de todas las carreras alcanza al 50% de la duración teórica de las mismas, llegando en algunas a superarlo.

Ahora bien, ¿a qué se hace referencia con el término rendimiento académico?

El rendimiento académico es un claro indicador del avance exitoso en la carrera de estudios de algún alumno en un momento particular, y a su vez también es un pronosticador de la posibilidad de completar exitosamente dicha carrera de estudios.

El término "rendimiento" tiene muchas implicancias, principalmente si se considera a las notas obtenidas por los alumnos como el referente casi exclusivo. Esta información puede generar, incluso una lectura ingenua, que centra sólo la responsabilidad académica en el alumno. Sin embargo, la responsabilidad institucional es clave para evaluar lo que se entiende por rendimiento. Más allá de las condiciones internas a las instituciones y de las prácticas docentes, resulta imprescindible también conocer las características que aportan quienes son los receptores de la labor docente. Esta información puede contribuir a estimar algunas de las razones que inciden en el rendimiento y la deserción de los alumnos universitarios (Toer, 2000).

Debe tenerse en cuenta que se trata de un constructo teórico complejo y multidimensional, atravesado y determinado por múltiples factores sociales, económicos, históricos, institucionales e individuales. Por tal motivo el rendimiento académico ha sido representado de diferentes maneras en los diversos estudios que han abordado el tema. En algunos, está representado sólo por el número de materias aprobadas por un alumno en una carrera, en otros por los resultados de tests específicamente diseñados o el promedio de notas de las asignaturas cursadas. Esta variedad de interpretaciones del concepto de rendimiento académico está ligada a las particularidades de las investigaciones en cuestión, principalmente al momento histórico en que se realiza la investigación y las concepciones de quienes llevan a cabo y financian la misma. Restringir el concepto a uno solo de estos indicadores, supone una postura ingenua y hasta simplista de lo que el acto educativo significa, pues equivale a descontextualizarlo, aislándolo de la realidad social e histórica de la que forma parte. En una mirada contextualizada, el rendimiento académico es el producto de condiciones institucionales (diseño curricular, práctica docente, valores y concepciones institucionales, etc.), socioeconómicas (situación laboral, estado civil, nivel educativo del grupo familiar, entre otras) e individuales (formación previa, hábitos de estudios, etc.) de los estudiantes.

Para la construcción de modelos de predicción del rendimiento académico de los estudiantes, se utilizará información referida a la situación socioeconómica y el desempeño académico en el período 2001- 2008 de los estudiantes de las diversas carreras de FACENA. Para el análisis de esta información se utilizarán técnicas de minería de datos.

La minería de datos se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en diferentes formatos. El objetivo es encontrar modelos inteligibles a partir de los datos, descubrir patrones cuya utilización apoye decisiones que reporten beneficios a la organización (Hernández Orallo et al, 2004).

Para cumplir sus objetivos, son dos los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, que proceden generalmente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos, etc.), y por el otro, usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos caso, la utilidad del conocimiento extraído está relacionada con la comprensibilidad del modelo inferido (Hernández Orallo et al, 2004).

Esta tecnología emergente combina el análisis estadísticos y la gestión de las bases de datos para extraer información desde los datos, y se presenta como un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras tecnologías (Thuraisingham, 2000).

Las técnicas que conforman el campo de la Minería de Datos buscan descubrir, en forma automática, el conocimiento contenido en la información almacenada en las bases de datos de las organizaciones. Por medio del análisis de datos, se pretende descubrir patrones, perfiles y tendencias. Es importante que estas técnicas sean las adecuadas al problema abordado. En este sentido, se pueden establecen dos grandes grupos de técnicas ó métodos analíticos: los métodos simbólicos y los métodos estadísticos (Britos, 2005).

Entre los métodos simbólicos se incluyen a las Redes Neuronales, Algoritmos Genéticos, Reglas de Asociación, Lógica Difusa, entre otros. Estos derivan del campo de la Inteligencia Artificial.

Los métodos estadísticos están constituidos por las técnicas del Análisis Multivariante de Datos, tales como Regresión Lineal simple y Múltiple, Regresión

No Lineal, Regresión Logística, Análisis Discriminante, Árboles de Regresión, entre otras. Las técnicas de esta categoría, de alguna manera, constituyen la piedra basal de la Minería de Datos (Britos, 2005).

En este estudio se utilizarán las siguientes metodologías para analizar las variables socioeconómicas relacionadas con el rendimiento académico: La Regresión Logística, los Arboles de Decisión y las Redes Neuronales.

El modelo de Regresión Logística es un método lineal que intenta modelizar la probabilidad de ocurrencia de un evento de interés. La variable dependiente es categórica dicotómica o policotómica, a los efectos de facilitar la interpretación (Britos, 2005).

Es una técnica adecuada cuando se pretende hacer una clasificación basada en las características de los datos. Una ventaja adicional de esta técnica es que no requiere la normalidad estricta de los datos, además muchos estudios han evidenciado otras características que hacen de la regresión logística una buena herramienta para la categorización (García Jiménez et al, 2000).

Las Redes Neuronales son modelos computacionales inspirados en las características neurofisiológicas del cerebro humano y están formadas por un gran número de neuronas dispuestas en varias capas e interconectadas entre sí mediante conexiones con pesos. Una neurona sobre un conjunto de nodos N es una tripleta (X, f, Y), donde X es un subconjunto de N, Y es un único nodo de N y f es una función neuronal que calcula un valor de salida para Y basado en una combinación de los valores de los componentes de X, es decir $y = f(\sum_{x_i \in N} w_i x_i)$. Los pesos wi pueden ser positivos o negativos, reproduciendo el carácter excitador o inhibitorio de la sinapsis de las neuronas. Las redes neuronales usan un proceso de aprendizaje por analogía donde los pesos de las conexiones son ajustados para reproducir un conjunto de datos representativo del problema a aprender. Las redes neuronales constituyen herramientas analíticas que permiten examinar los datos con el objeto de descubrir y modelar las relaciones funcionales existentes entre las variables. Pueden comportarse como técnicas de aproximación o de clasificación universales (Castillo et al, 1999).

Como antecedentes de aplicación de la técnica de redes neuronales en el ámbito de educación pueden mencionarse los trabajos de González (1999), Salgueiro et al (2006), Borracci y Arribalzaga (2005).

Los árboles de decisión son una serie de decisiones o condiciones organizadas de forma jerárquica, a modo de árbol. Son muy útiles para encontrar estructuras en espacios de alta dimensionalidad y en problemas que mezclan datos categóricos y numéricos.

Básicamente, un árbol de decisión es un árbol donde cada nodo representa una condición o test sobre algún atributo y cada rama que parte de ese nodo corresponde a un posible valor para ese atributo. Finalmente, las hojas representan el valor de la variable predicha. Esta técnica se usa en tareas de clasificación, agrupamiento y regresión. Los árboles de decisión que se usan para predecir variables categóricas se llaman árboles de clasificación, mientras que los árboles de decisión que se utilizan para predecir variables continuas se llaman árboles de regresión (Alcover et al, 2007).

Como antecedente de aplicación de la técnica de árboles de regresión aplicado el rendimiento de alumnos universitarios puede mencionarse aBacallao Gallestey et al (2004).

2. LINEAS DE INVESTIGACION y DESARROLLO

a) **Preprocesamiento de los datos: Actualización de la base de datos oportunamente diseñada dentro del marco del proyecto PI 005/06, con los datos socioeconómicos y del estado académico de los alumnos de las carreras de FACENA a diciembre de 2008**, a fin de proveer la información que constituye el soporte de las actividades del proyecto. Esto implica incorporar los datos del estado académico de los alumnos, que provee el Departamento Estudios de la FACENA, y los datos que surgen del formulario de ingreso que completan los alumnos al momento de ingresar a la UNNE.

b) **Modelado de datos y predicción del rendimiento académico en el primer cuatrimestre del primer año de estudios en función del nivel de conocimientos matemáticos previos de los ingresantes a la FACENA:** Los datos del diagnóstico se utilizarán como variables explicativas en un modelo que permita predecir el rendimiento de los alumnos al finalizar el primer cuatrimestre, empleando métodos simbólicos y estadísticos.

c) **Formulación y ajuste de modelos para predecir el rendimiento académico de los alumnos en función de las características socioeconómicas de los mismos,** empleando métodos estadísticos y simbólicos. Numerosas investigaciones han encontrado vinculación entre las condiciones socioeconómicas y personales de un individuo (tales como, edad, sexo, lugar de procedencia, etc.), y su rendimiento académico. Para el análisis de esta relación se requiere, previamente, una etapa de preprocesamiento de los datos, que comprende los siguientes pasos: Integración, Reconocimiento y Limpieza, Transformación y Reducción. Esta etapa tiene por objeto mejorar

la calidad de los datos, teniendo en cuenta que, a lo largo del período de estudio, se ha modificado, en más de una oportunidad, el diseño del formulario de ingreso, situación que exige un análisis detallado para la determinación de equivalencias entre los distintos valores de las variables en estudio.

Para la construcción y ajuste de los modelos de predicción del rendimiento académico de los alumnos se utilizarán métodos de minería de datos simbólicos y estadísticos, previéndose realizar un estudio comparativo entre ambos grupos de metodologías, con el objeto de contrastar el desempeño y la eficiencia de las mismas en el problema de la predicción del rendimiento académico de los estudiantes.

3. RESULTADOS OBTENIDOS/ESPERADOS

Dentro de las líneas de trabajo mencionadas, se han obtenido los siguientes resultados:

a) Se han estudiado técnicas de preprocesado de datos para mejorar la calidad de la información obtenida desde los sistemas de información existentes, y para mantener actualizado un repositorio con toda la información sistematizada existente en la unidad académica respecto del desempeño de los alumnos (Dapozo et al, 2007).

b) Se ha analizado el perfil socioeconómico y educativo de los alumnos ingresantes de la FACENA y su relación con su rendimiento académico, medido en términos de su desempeño en la primera asignatura de Matemática, en el primer año de carrera universitaria, utilizando técnicas clásicas de minería de datos (Porcel et al, 2008).

c) Se ha analizado el rendimiento académico de los alumnos de las trece carreras de grado de la FACENA, utilizando indicadores basados en la relación entre el número de exámenes rendidos y el número de asignaturas aprobadas por los mismos, estimados mediante regresión lineal paramétrica y no paramétrica. Se trazaron gráficos de dispersión para cada carrera que permitieron observar la eficiencia de los alumnos en los exámenes, y se calculó además una matriz de correlación de los indicadores. (Porcel et al, 2009)

Como resultados esperados, se espera poder predecir el rendimiento académico de los estudiantes de FACENA, para lo cual se ha planificado la construcción de modelos cuantitativos predictivos del desempeño estudiantil en base a la información disponible sobre las condiciones socioeconómicas de los estudiantes, a sus conocimientos matemáticos previos y los datos aportados por los mismos alumnos a través de instrumentos diseñados especialmente.

Los resultados de la investigación aportarán a un mayor conocimiento de los posibles factores que inciden en el desempeño de los alumnos. Esta información permitirá que, desde la gestión institucional, se aborden mecanismos correctivos o superadores que contribuyan al mejoramiento de los índices de desgranamiento, abandono, bajo rendimiento académico y prolongación excesiva de la duración de la carrera, comunes en el ambiente aniversario nacional, principalmente notorios en la actuación de los alumnos en el primer año de carrera.

Por otra parte, en el marco actual de las universidades nacionales, los procesos de evaluación y acreditación de títulos se basan en la construcción de indicadores que permiten descubrir fortalezas y debilidades de la formación universitaria, por lo que "reflexionar sobre todos los elementos que la evaluación del rendimiento del alumnado proporciona se convierte en un mecanismo claro para la mejora de la calidad del proceso educativo" (Muñoz, 2005).

Finalmente, a nivel de las políticas educativas nacionales, la formación en Informática, Ingeniería y Ciencias Básicas, se considera prioritaria. Se destinan recursos especiales como becas, planes de tutorías y otras acciones tendientes a promover titulaciones en estas áreas, así como también, mejorar el índice de graduados y de retención de alumnos. Conocer las causas que subyacen en el rendimiento académico, permitirá mejorar estas iniciativas por el aporte de mayor información.

4. FORMACION DE RECURSOS HUMANOS

En este proyecto se enfatiza el enfoque interdisciplinario dado que sus integrantes proceden de distintas disciplinas: Matemática, Estadística e Informática. Esto permite un abordaje sistémico de los problemas de la investigación, a la vez que se complementan y enriquecen las distintas miradas disciplinares. Los alumnos en proceso de formación de la Licenciatura en Matemática y la Licenciatura en Sistemas de Información que se suman al proyecto en calidad de becarios o para la realización del Trabajo Final de Aplicación, requisito académico de la carrera de Sistemas, tienen la oportunidad de aplicar y ampliar sus conocimientos en estas aplicaciones interdisciplinarias que posibilita la Minería de Datos.

5. BIBLIOGRAFIA

- ALCOVER, R., BENLLOCH J., BLESA, P., CALDUCH, M., CELMA, M., FERRI C., HERNÁNDEZ ORALLO, J., Y OTROS "Análisis del rendimiento académico en los estudios de Informática de la Universidad Politécnica de Valencia aplicando técnicas de Minería de Datos". XII Jornadas de Enseñanza Universitaria de la Informática 2007. Disponible En http://bioinfo.uib.es/~joemiro/aenui/procjenui/jen2007/alanal.pdf
- BACALLAO GALLESTEY C., PARAPAR DE LA RISTRA, J., ROQUE GIL M., BACALLOA GUERRA J. "Arboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico". Revista Cubana de Educación Médica Superior, vol.18, N°3. 2004. Disponible en: http://bvs.sld.cu/revistas/ems/vol18_3_04/ems02304.htm
- BORRACCI, R. A., ARRIBALZAGA, E. B. "Aplicación de análisis de conglomerados y redes neuronales artificiales para la clasificación y selección de candidatos a residencias médicas". Educación Médica, Vol 8, Nº 1. ISSN 1575-1813. Barcelona. 2005.
- BRITOS, P. V. "Minería de Datos". Buenos Aires: Nueva Librería. 2005.
- CASTILLO, E., COBO, A., GUTIÉRREZ, J. M., PRUNEDA, R. E. "Introducción a las Redes Funcionales con Aplicaciones. Un Nuevo Paradigma Neuronal". Editorial Paraninfo S.A. Madrid. España. 1999.
- DAPOZO, G., PORCEL, E., LÓPEZ, M. V.; BOGADO, V. "Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios". IX Workshop de Investigadores en Ciencias de la Computación (WICC 2007). Trelew. Chubut. Argentina. 2007.
- GARCÍA JIMÉNEZ, M. V., ALVARADO IZQUIERDO, J. M. y JIMÉNEZ BLANCO, A. "La predicción del rendimiento académico: regresión lineal versus regresión logística". Psicothema, 12 (2), 248-252. 2000. Disponible en http://redalyc.uaemex.mx/redalyc/pdf/727/72797059.pdf.
- GONZÁLEZ, D.S. "Detección de alumnos de riesgo y medición de la eficiencia de centros escolares mediante redes neuronales". Biblioteca de Económicas y Empresariales. Servicios de Internet. Universidad Complutense de Madrid. 1999.
- HERNÁNDEZ ORALLO, J., FERRI RAMÍREZ, C. y RAMÍREZ QUINTANA M. J. "Introducción a la Minería de Datos". España: Prentice Hall. Pearson Education. 2004.
- MUÑOZ, S. "Indicadores de rendimiento académico del alumnado de la universidad de La Laguna. Jornadas sobre Políticas de Calidad en la Universidad de La Laguna", 2005.
- PORCEL E. A., DAPOZO GLADYS N., LÓPEZ M. "Técnicas clásicas de minería de datos aplicadas al estudio del rendimiento académico de alumnos de primer año de carreras de la FACENA". Comunicaciones Científicas y Tecnológicas 2008. Universidad Nacional del Nordeste. Corrientes. Argentina. 2008.
- PORCEL, E., LÓPEZ, M. V., DAPOZO, G., CAPUTO, L. "Relación entre el número de exámenes rendidos y el número de asignaturas aprobadas como indicador del rendimiento académico de alumnos universitarios". XXII Encuentro Nacional de Docentes de Investigación Operativa (ENDIO). XX Escuela de Perfeccionamiento en Investigación Operativa (EPIO). Buenos Aires. Argentina. 2009.
- SALGUEIRO, F., COSTA, G., CÁNEPA, S., LAGE,F., KRAUS, G., FIGUEROA, N., CATALDI, Z. "Redes Neuronales para predecir la aptitud del alumno y sugerir acciones". VIII Workshop de Investigadores en Ciencias de la Computación (WICC 2006). Buenos Aires. Argentina. 2006.
- THURAISINGHAM, B. "A primer for understanding and applying Datamining". IT Professional, 2 (1), 28-31. 2000.
- TOER, M. "El caso de los ingresantes de 1998 al Ciclo Básico Común de la Universidad de Buenos Aires, para seguir carreras de la Universidad de Derecho, Ciencias Económicas y Ciencias Sociales". Buenos Aires, Argentina: Instituto de Investigaciones Gino Germani, FCSoc., Ciclo Básico Común, Universidad de Buenos Aires. 2000. Disponible en http://caraya.cbc.uba.ar/dat/sbe/perfil/perfil.html#1.